AFRL-SR-AR-TR-07-0390

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	9/10/07	Final Report: 3/01/04 – 7/31/07

4. TITLE AND SUBTITLE
Constraint-based Integration of Geospatial and Online Sources

5. FUNDING NUMBERS
FA9550-04-1-0105

6. AUTHORS
Craig A. Knoblock

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Southern California / Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 North Randolph Street, Suite 325
Arlington VA 22203-1768

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12A. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12B. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 words)***

In this research project, we developed a constraint-based approach to integrating traditional and non-traditional online geographic data sources. There were three areas where we made significant advances. First, we developed a constraint satisfaction framework to integrate data sources for the labeling of buildings in satellite imagery. Second, we developed an automatic approach for the integration of maps, vector data, and high-resolution satellite imagery. Third, we developed an approach to automatically extract the road network and the textual labels from a raster map.

14. SUBJECT TERMS
Constraint programming, geographic data integration, information integration, conflation, satellite imagery

15. NUMBER OF PAGES
12 pages

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
289-102

2007 Final Performance Report

Constraint-based Integration of Geospatial and Online Sources

USAF, Air Force Office of Scientific Research
Award Number: FA9550-04-1-0105
Period of Performance: 3/1/04 – 7/31/07

Craig A. Knoblock (PI)
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
Phone: 310-448-8786
Fax: 310-822-0751
knoblock@isi.edu

September 10, 2007

Status of the Effort:

This research project addressed the problem of using a constraint-based approach to integrating traditional and non-traditional online geographic data sources. In this work, we made three significant advances. First, we developed a constraint satisfaction framework to integrate data sources for the labeling of buildings in satellite imagery. Second, we developed an automatic approach for the integration of maps, vector data, and high-resolution satellite imagery. Third, we developed an approach to automatically extract the road network and the textual labels from a raster map.

Accomplishments/New Findings:

In this section we describe our contribution on building identification in satellite imagery, extraction of road networks and textual labels from raster maps, and integration and alignment of maps, vector data, and satellite imagery. This project resulted in one Ph.D. and three Masters theses on these topics, which are available from http://www.isi.edu/~knoblock:

- Rahul Bakshi
  Integration and reasoning about online sources to accurately geocode addresses.
  Master's thesis, University of Southern California, 2004.

- Matthew Michelson
  Building Queryable Datasets from Ungrammatical and Unstructured Sources.
  Master's thesis, University of Southern California, 2005.

- Kenneth M. Bayer
  Reformulating Constraint Satisfaction Problems with Application to Geospatial Reasoning.
  Master's thesis, University of Nebraska-Lincoln, 2007.

- Ching-Chien Chen.
  Automatically and Accurately Conflating Road Vector Data, Street Maps and Orthoimagery.
  PhD thesis, University of Southern California, 2005.

Building Identification in Satellite Imagery

Our solution to the building identification problem combines known addressing characteristics seen throughout the world with the integration of information from publicly available data sources. However, the integration of data sources is a non-trivial task. As such, we developed an approach that uses Constraint Processing (CP) techniques to associate addresses with the buildings in a satellite image using publicly available data sources such as a phone book (Michalowski and Knoblock, 2005). We proposed a constraint model of the problem and used an existing solver called CPlan to find all

possible matches of addresses to buildings that are consistent with a phone book and with the geographical layout in the image. This work established the feasibility of the approach and identified an important new area where CP techniques are useful for solving real-world problems. However, the scalability of our proposed techniques required refinements to our initial approach. As such, we improved our initial constraint model and developed a customized constraint solver that leverages the structure of the building identification problem to improve scalability.

Improved Constraint Model

We improved the constraint model by reducing the number of variables and the arity of some of the constraints. We did this by integrating redundant representations of values as intervals and discrete sets in order to speed up constraint propagation. Our new solver (discussed below) uses the representations selectively and ensures that the values are consistent at any point in time. Further, the new constraint model provides a mechanism that can be switched on and off to exploit common features of real-world situations, such as the existence of known landmarks in the map and numbering schemas across the world. Such an approach allows us to exploit additional information available for a given area. This additional information is used to further constrain our problem and our empirical evaluation shows that these further constrained problems lead to higher solution quality and shorter runtime.

Customized Constraint Solver

To further improve the scalability of our building identification problem-solving approach, we developed a customized solver (Bayer et al. 2007) to replace the generic one used initially. This customized solver includes many improvements over the previously used one. This solver is based on backtrack search and exploits structural properties of a problem instance, such as identifying backdoor variables and exploiting them to decompose the problem into tractable components. It also uses four reformulation techniques to reduce the cost of problem solving. These techniques are (1) reformulating the building identification problem from a counting problem to a satisfiability one, (2) reducing the domains size of variables in the scope of a global constraint that we identify and characterize, (3) relaxing the satisfiability problem into a matching problem, and (4) using symmetry to generate efficiently all possible solutions of the relaxed version of the original building identification counting problem.

Finally, to further improve the scalability of our problem-solving approach, we revisited the task definition posed in our previous work and reformulated the query from the expensive task of "finding all consistent *combinations* of addresses" to the significantly cheaper one of "finding all possible addresses for each building". Both queries yield exactly the same results but the former is much more computationally expensive. The computational savings are achieved by solving many small sub problems and combining the solutions rather then solving one very large instance of the problem as a whole. These reformulated queries are supported by our customized solver. Both of these assertions have been validated by our empirical evaluation. Subsequently, this allows us to tackle larger problem instances than was previously possible, making the evaluation of problem-solving performance on real-world problems feasible.

Automatic Extraction of Road Labels from Raster Maps

In order to support the automatic labeling of buildings in images, we need to know the names of the roads on which they are located. Raster maps contain the labels for roads. We developed an approach to automatically extract road intersections from maps and then automatically align the maps with satellite imagery. Raster maps typically contain multiple layers that represent roads, symbols, etc. The road layer needs to be automatically separated from other layers before road intersections can be extracted. The steps of our approach are illustrated in Figure 1. We combine a variety of image processing and graphics recognition methods to automatically eliminate the other layers and then extract the road intersection points. During the extraction process, we determine the intersection connectivity (i.e., number of roads that meet at an intersection) and the road orientations. This information helps in matching the extracted intersections with intersections from known sources.

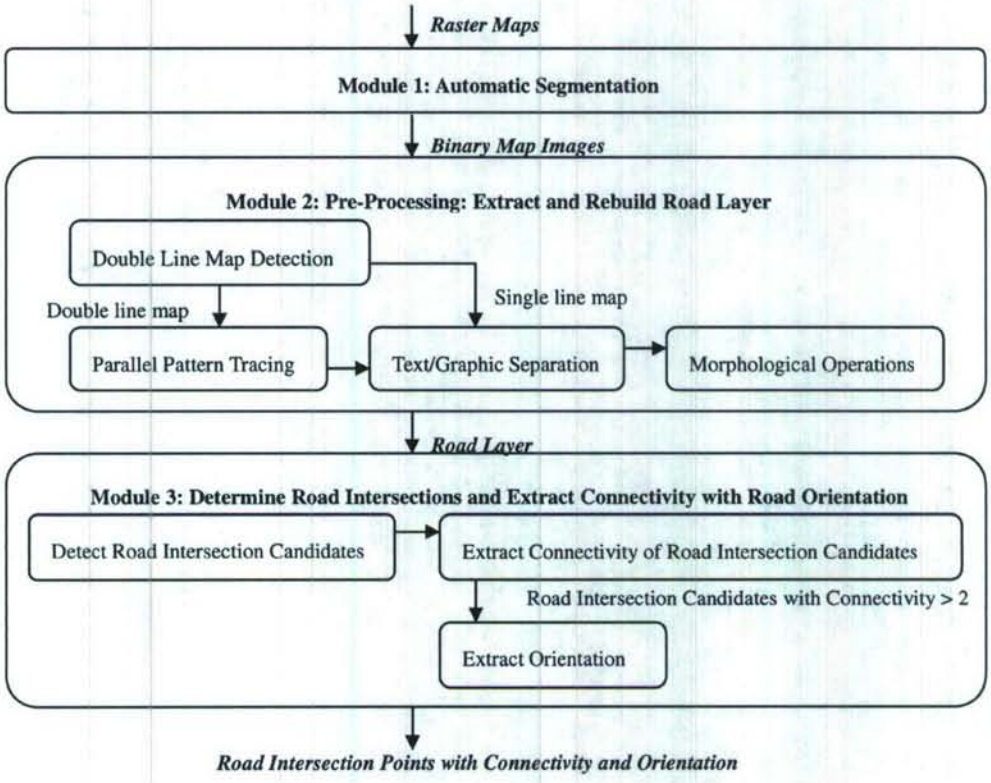

Figure 1. Road Intersection Extraction Process

We applied the techniques to a set of 48 randomly selected raster maps and achieved over 90% precision with over 75% recall. The results help the conflation system to determine the precise coverage and scale of the raster maps (Chen 2005). In addition, we applied

our technique on randomly returned maps from image search engines and successfully identify the road intersection points for conflation systems to identify the geocoordinates (Desai et al. 2005).

We also developed an approach to automatically extract both the road names and the road network from the raster maps. We designed an algorithm that uses the 2-D discrete cosine transformation (DCT) coefficients and support vector machines (SVM) to automatically classify pixels on raster maps into line or character classes. The classification results (i.e., line and character image) can be further used in vectorization components and OCR components to pull out the information such as the geometries and names of streets from the raster maps.

DCT has played an important role in many texture classification applications for its outstanding ability to generate distinct features for different texture representations. It transforms an image into the frequency domain where the strength of each frequency is represented by one of the DCT coefficients. Within a local area (i.e., a DCT window), the textures of the foreground and the background are different since the colors of the background are consistent while the colors of the foreground change frequently. Among the foreground objects, lines and characters also have different texture.

In our algorithm, the classification is pixel-based. Initially every pixel is automatically classified into background or foreground classes using a threshold. The foreground pixels alone are sent to the SVM for line or character pixel classification. Support Vector Machines are widely used in many research fields that require classification, especially in the area of pattern recognition. In the training process of our algorithm, the SVM constructs hyperplanes in a multidimensional space (i.e., the feature space of the DCT coefficients) that separates pixels of different classes (i.e., line and character classes). SVM can quickly generate a model from a small set of training data and it is robust to noisy data.

Integrating Maps and Imagery

There are maps available for locations throughout the world. However, for many of these maps, the geocoordinates and scale of the maps are unknown. Even if this information is known, accurately integrating maps and imagery from different data sources remains a challenging task. This is because spatial data obtained from various data sources may have different projections and different accuracy levels. If the geographic projections of these datasets are known, then they can be converted to the same geographic projections. However, the geographic projection for a wide variety of geo-spatial data available on the Internet is not known. To address this problem, we built on our previous work on automatic vector to image conflation and developed efficient techniques to the problem of automatically conflating maps with satellite imagery (Chen et al. 2003a; Chen et al. 2004a).

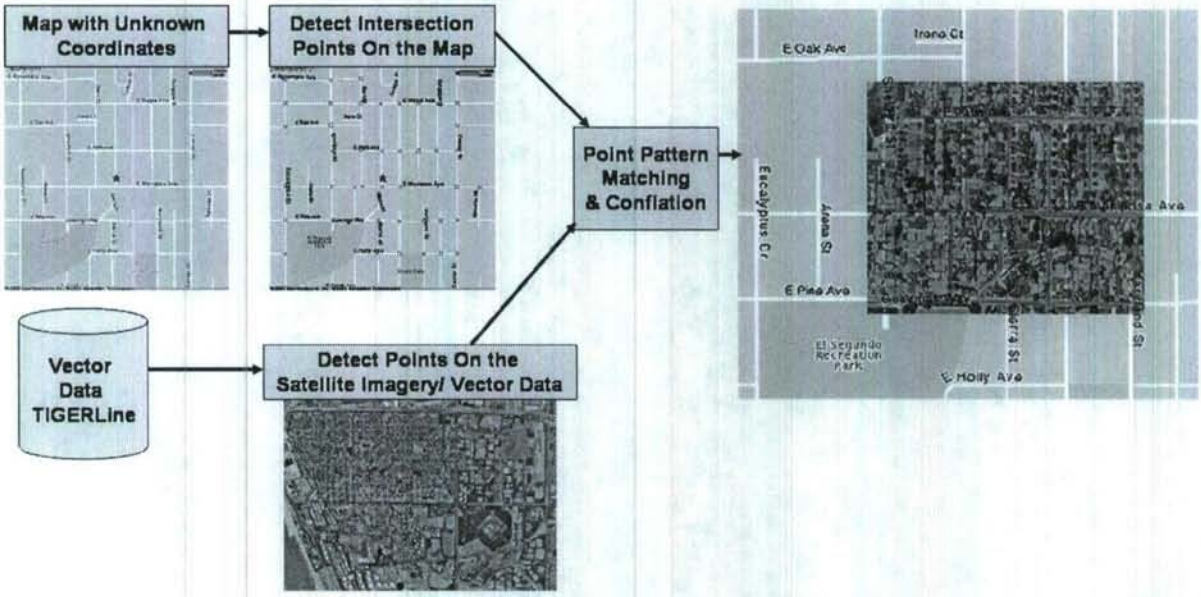

Figure 2: Automatic conflation of maps with imagery

The steps of our approach are illustrated in Figure 2. First, we utilize the techniques described above to align the road vector data with the imagery to identify the intersection points on the imagery. Then we apply techniques for identifying the intersections on maps (Sebok et al. 1981; Musavi et al. 1988), which we have extended to support maps with double-lined roads and maps with lots of extraneous data such as on topographic maps. Next, we apply a specialized point matching algorithm (Irani et al. 1999) to compute the alignment between the two sets of intersection points. This matching problem is challenging because of the potential of both missing and extraneous intersection points from the map intersection detection algorithms. Finally, we use the resulting set of control point pairs to automatically conflate the map and image.

Experimental results on the city of El Segundo demonstrate that our approach leads to remarkably accurate alignments of maps and satellite imagery. The aligned map and satellite imagery supports inferences that could not have been made from the map or imagery alone.

List of Personnel Associated with the Research Effort:

Craig Knoblock, PI
Maria Muslea, Research Scientist
Martin Michalowski, Graduate Research Assistant
Rahul Bakshi, Graduate Research Assistant
Snehal Thakkar, Graduate Research Assistant
Ching-Chien Chen, Graduate Research Assistant
Yao-Yi Chiang, Research Scientist
Sneha Desai, Graduate Research Assistant

Cyrus Shahabi, Associate Professor, USC
Berthe Choueiry, Associate Professor, University of Nebraska
Ken Bayer, Research Assistant, University of Nebraska

Publications:

Rahul Bakshi, Craig A. Knoblock, and Snehal Thakkar.
Exploiting online sources to accurately geocode addresses.
In Proceedings of the 12th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS'04), 2004.

Rahul Bakshi.
Integration and reasoning about online sources to accurately geocode addresses.
Master's thesis, University of Southern California, July, 2004.

Ching-Chien Chen, Cyrus Shahabi, and Craig A. Knoblock.
Utilizing road network data for automatic identification of road intersections from high resolution color orthoimagery.
In Proceedings of the 2nd Workshop on Spatio-Temporal Database Management - STDBM'04, 2004.

Ching-Chien Chen, Craig A. Knoblock, Cyrus Shahabi, Snehal Thakkar, and Yao-Yi Chiang.
Automatically and accurately conflating orthoimagery and street maps.
In Proceedings of the 12th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS'04), 2004.

Martin Michalowski and Craig A. Knoblock.
A Constraint Satisfaction Approach to Geospatial Reasoning,
In Proceedings of The Twentieth National Conference on Artificial Intelligence (AAAI-05), Pittsburgh PA, July 2005.

Martin Michalowski, Snehal Thakkar, and Craig A. Knoblock.
Automatically Utilizing Secondary Sources to Align Information Across Sources,
AI Magazine, Special Issue on Semantic Integration, Vol. 26, No. 1, pp. 33-45, Spring 2005.

Ching-Chien Chen.
Automatically and Accurately Conflating Road Vector Data, Street Maps and Orthoimagery.
PhD thesis, University of Southern California, 2005.

Sneha Desai, Craig A. Knoblock, Yao-Yi Chiang, Kandarp. Desai, and Ching-Chien Chen.
Automatically filtering and categorizing street maps on the web.
In The 2nd International Workshop on Geographic Information Retrieval (GIR'05), 2005.

Greg Barish and Craig A. Knoblock.
An expressive and efficient language for software agents.
Journal of Artificial Intelligence Research, 23:625—666, 2005.

Yao-Yi Chiang, Craig A. Knoblock, and Ching-Chien Chen.
Automatic extraction of road intersections from raster maps.
In The 13th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS'05), Bremen, Germany, November, 2005.

Snehal Thakkar, Jose Luis Ambite, and Craig A. Knoblock.
Composing, optimizing, and executing plans for bioinformatics web services.
VLDB Journal, Special Issue on Data Management, Analysis and Mining for Life Sciences, 14(3):330--353, Sep 2005.

Mark James Carman and Craig A. Knoblock.
Inducing source descriptions for automated web service composition.
In Proceedings of the AAAI 2005 Workshop on Exploring Planning and Scheduling for Web Services, Grid, and Autonomic Computing, Technical Report WS-05-03. AAAI Press, 2005.

Craig A. Knoblock, Pedro Szekely, and Rattapoom Tuchinda.
A mixed-initiative system for building mixed-initiative systems.
In Proceedings of the AAAI Fall Symposium on Mixed-Initiative Problem-Solving Assistants, 2005.

Ion Muslea, Steve Minton, and Craig A. Knoblock.
Active learning with multiple views,
Journal of Artificial Intelligence Research, 27:203—233, 2006.

Ching-Chien Chen, Craig A. Knoblock, and Cyrus Shahabi.
Automatically conflating road vector data with orthoimagery,
Geoinformatica, 10(4):495—530, 2006.

Yao-Yi Chiang and Craig A. Knoblock.
Classification of Line and Character Pixels on Raster Maps Using Discrete Cosine Transformation Coefficients and Support Vector Machines
International Conference on Pattern Recognition (ICPR 2006), 2006.

Kristina Lerman, Anon Plangrasopchok, and Craig Knoblock.
Automatically labeling the inputs and outputs of web services.
In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), Boston, MA, 2006.

Mark James Carman and Craig A. Knoblock.
Learning semantic descriptions of web information sources,

In Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI), 2007.

Rattapoom Tuchinda, Pedro Szezly, and Craig A. Knoblock.
Building data integration queries by demonstration,
In IUI '07: Proceedings of the international conference on Intelligent user interface, 2007.

Martin Michalowski, Craig A. Knoblock and Berthe Y. Choueiry.
Exploiting Problem Data to Enrich Models of Constraint Problems,
In the Proceedings of 6th International Workshop On Constraint Modeling and Reformulation (ModRef'07), 2007

Kenneth M. Bayer, Martin Michalowski, Berthe Y. Choueiry and Craig A. Knoblock.
Reformulating CSPs for Scalability with Application to Geospatial Reasoning,
In Proceedings of the 13th International Conference on Principles and Practice of Constraint Programming (CP-07), 2007

Martin Michalowski, Craig A. Knoblock, and Berthe Y. Choueiry.
Reformulating Constraint Models Using Input Data,
In Proceedings of the 7th Symposium on Abstraction, Reformulation and Approximation (SARA-07), Research Summary, 2007

Kenneth M. Bayer, Martin Michalowski, Berthe Y. Choueiry, and Craig A. Knoblock.
Reformulating Constraint Satisfaction Problems to Improve Scalability,
In Proceedings of the 7th Symposium on Abstraction, Reformulation and Approximation (SARA-07), 2007

Matthew Michelson and Craig A. Knoblock,
Mining Heterogeneous Transformations for Record Linkage,
In Proceedings of the 6th International Workshop on Information Integration on the Web (IIWEB-07), 2007

Matthew Michelson and Craig A. Knoblock,
Beginning to Understand Unstructured, Ungrammatical Text: An Information Integration Approach,
In Proceedings of the AAAI Spring Symposium on Machine Reading, 2007

Matthew Michelson and Craig A. Knoblock,
An Automatic Approach to Semantic Annotation of Unstructured, Ungrammatical Sources: A First Look,
In Proceedings of the 1st IJCAI Workshop on Analytics for Noisy Unstructured Text Data (AND-07), 2007

Kristina Lerman, Anon Plangrasopchok, and Craig A. Knoblock.
Semantic labeling of online information sources,
International Journal on Semantic Web and Information Systems, 3(3): 36—56, 2007.

Matthew Michelson and Craig A. Knoblock,
Unsupervised Information Extraction from Unstructured, Ungrammatical Data Sources on the World Wide Web,
International Journal of Document Analysis and Recognition (IJDAR), Special Issue on Analytics for Noisy Unstructured Text Data, (To appear)

Consultative and Advisory Functions

Presentation of research on this project to:

- Caryn Bain, Program Manager for SOCOM Advanced Concept Technology Demonstration for Psychological Operations
- Dr. Michael Macedonia, Chief Scientist and Technical Director for The Simulation, Training, and Instrumentation Command (STRICOM)
- Mike Full, NTA Program Manager for the National Geospatial Intelligence Agency

Interactions/Transitions:

Presentation on Geospatial Data Integration at AFRL in Rome, NY in May, 2004. Visit was hosted by John Salerno.

Invited to give presentations on this work to the NSA, CIA, and NGA on Oct 19-20 by

- W. Arnold Landvoigt, Office of Tradecraft for Analysis / Advanced Analysis Lab National Security Agency

Presentation on Geospatial Data Integration at NGA in St. Louis on August 24, 2005.

Presentation on Geospatial Data Integration to the Army Topographic Engineering Center at Ft. Belvoir in VA on July 21, 2005.

Talk on Geospatial Data Integration at the CAL IT2 seminar series at UC Irvine on July 28, 2005.

Presentation by Craig Knoblock at the AFOSR PI Meeting in Florence, Italy in July, 2006.

Presentation by Craig Knoblock at AFOSR in Washington, DC in May, 2006.

Presentation by Yao-Yi Chiang at the International Symposium on Advances in Geographic Information Systems in Bremen, Germany in November, 2005.

Presentation by Yao-Yi Chiang at the International Conference on Pattern Recognition in Hong Kong, August 2006.

Presentation by Kristina Lerman at the National Conference on Artificial Intelligence (AAAI-06) in Boston, July 2006.

Presentation by Craig Knoblock at the Twentieth International Joint Conference on Artificial Intelligence in January 2007.

Presentation by Rattapoom Tuchinda at the International Conference on Intelligent User Interfaces in 2007.

Presentation by Martin Michalowski at the 7th Symposium on Abstraction, Reformulation and Approximation in Whistler, July 2007.

Presentation by Berthe Choueiry at the 7th Symposium on Abstraction, Reformulation and Approximation in Whistler, July 2007.

Presentation by Matthew Michelson at the 6th International Workshop on Information Integration on the Web in Vancouver, July 2007.

Presentation by Matthew Michelson at the AAAI Spring Symposium on Machine Reading in June 2007.

Presentation by Craig Knoblock at the 1st IJCAI Workshop on Analytics for Noisy Unstructured Text Data in January 2007.

Technology on map and image fusion has been licensed to Geosemble Technologies for commercialization.

Discoveries/Inventions/Patent Disclosures

- Invention disclosure and patent application filed on Automatic Vector to Imagery and Map to Imagery Registration
- Patent pending on a US Utility patent application on Automatically and Accurately Conflating Road Vector Data, Street Maps, and Orthoimagery, Serial #11/169,076, Filing Date: 06/28/2005

Honors/Awards

- Craig Knoblock was named a Fellow of the American Association of Artificial Intelligence in 2004
- Craig Knoblock was elected President-elect of the ICAPS Council, 2004
- Craig Knoblock gave an invited talk at the International Conference on Case-Based Reasoning on August 25, 2005. The topic of the talk was on Learning to Optimize Plan Execution in Information Agents, which was the topic of our previous AFOSR grant.
- Craig Knoblock gave an invited talk on Geospatial Data Integration at the University of Trento in Trento, Italy on July 19, 2006.
- Craig Knoblock was selected as Conference Chair for IJCAI 2011, 2007.